基于 BLSTM_Attention_CRF 模型的新能源汽车领域术语抽取

马建红[1]，张亚梅[1]，姚　爽[2]，张炳斐[1]，郭昌宏[1]

(1. 河北工业大学 计算机科学与软件学院，天津 300401; 2. 河北工业大学，天津 300401)

摘　要：为提高新能源汽车领域术语抽取准确率，面向新能源汽车专利文本提出一种领域术语抽取模型。传统的领域术语抽取方法过度依赖人工定义特征和领域知识，无法自动挖掘隐含特征，其识别性能过度依赖所选特征的质量。因此，从深度学习的角度出发，提出了一种基于 Attention 的双向长短时记忆网络（bidirectional long short-term memory, BLSTM）与条件随机场（conditional random fields，CRF）相结合的领域术语抽取模型（BLSTM_Attention_CRF 模型），并使用基于词典与规则相结合的方法对结果进行校正，准确率可达到 86%以上，该方法切实可行。

关键词：领域术语抽取；Attention 机制；双向长短时记忆网络；条件随机场；词典；规则

中图分类号：TP391　　doi: 10.3969/j.issn.1001-3695.2017.11.0741

Terminology extraction for new energy vehicle based on BLSTM_Attention_CRF model

Ma Jianhong[1], Zhang Yamei[1], Yao Shuang[2], Zhang Bingfei[1], Guo Changhong[1]

(1. *School of Computer Science & Software Hebei University of Technology*, *Tianjin* 300401, *China*; 2. *Hebei University of Technology*, *Tianjin* 300401, *China*)

Abstract: In order to improve the accuracy and recall rate of terminology extraction results in the field of new energy vehicles, this paper presented a domain terminology extraction model for the new energy vehicles patent text. Traditional domain terminology extraction methods rely too much on human-defined features and specialized domain knowledge to automatically mine implicit features whose recognition performance greatly depends on the quality of the selected features. In order to solve the problems, this paper proposed a model from the perspective of deep learning. First it extracted the domain terms by a combination of BLSTM（bidirectional long short-term memory, BLSTM）model based on the attention mechanism and CRF（conditional random fields, CRF）model(BLSTM_Attention_CRF model) , and then it corrected the result by a combination of dictionary and rules. Experimental results show that the accuracy of BLSTM-ATT-CRF model can reach more than 86%, which shows that the BLSTM-ATT-CRF model is effective to term extraction of new energy vehicles.

Key words: domain term extraction; attention mechanism; bidirectional long short-term memory; conditional random fields; dictionary; rules

0　引言

领域术语是以语音或文字为载体来表达或限定专业概念的约定性符号[1]，可以是词，也可以是词组，在我国又称为名词或科技名词。领域术语抽取技术在自然语言处理领域被广泛研究，并应用于多个领域,如文本分类、句法分析、自然语言生成、语料库语言学、统计机器翻译、信息检索、自动问答系统等领域[2]。随着科学技术的不断发展、新技术的不断涌现，以及互联网大数据、云计算时代的到来，使得特定领域的术语抽取需求不断扩大、更新，以往靠人工收集和非监督学习算法的抽取已经远远不能满足人们的需求，利用计算机自动抽取领域术语已经成为必然[3]。

专利文献具有新颖性、可靠性和权威性，是科技信息工作的重要研究对象，通常被认为是一种重要的知识来源。专利中的领域术语能够准确快捷地了解专利的方向以及核心技术，专利的有效利用能够提高国家和企业的发展速度[4~6]。由此，本文面向新能源汽车领域的专利文本抽取领域术语，基于深度学习建立自动抽取模型。经过大量分析专利文本及新能源汽车相关文献，发现专利文本中的新能源汽车领域术语主要存在以下特点：

收稿日期：2017-11-15；**修回日期：**2018-01-20

作者简介：马建红，女（1965-），教授，博士，主要研究方向为计算机辅助创新设计过程与方法、TRIZ、软件工程、CAI 软件技术（m_zh2002@126.com）；张亚梅，女（1991-），硕士研究生，主要研究方向为计算机辅助创新设计软件、软件工程、自然语言处理；姚爽，女（1987-），助理研究员，硕士，主要研究方向为自然语言处理；张炳斐，男（1993-），硕士研究生，主要研究方向为计算机辅助创新设计软件、软件工程、自然语言处理；郭昌宏，男（1993-），硕士研究生，主要研究方向为计算机辅助创新设计软件、软件工程、自然语言处理.

a）中文领域术语是一个开放的集合，随着时间转移会不断出现新词，所以抽取过程中的新词发现情况无法很好处理。

b）新能源汽车领域术语组合方式多变，词长主要从 2～10 字不等，其中包含较多的长术语和中英文混合的术语，如 AC/DC 电源、CAN 总线接口。

c）新能源汽车领域术语大多为嵌套和复合结构，如机油油量报警传感器，其中机油油量、传感器本身又是领域术语。据统计，新能源汽车专利文本中复杂术语的数量约占到 83%。

目前，众多相关学者对特定领域术语的抽取做了大量工作，主要有基于语言学规则的方法、基于统计的方法和两者结合起来使用的方法。周浪等人[7]根据术语的构词规律提出了构词法识别候选术语。基于语言学的方法是预先定义好许多规则模板，然后与待测语料进行模板匹配。其缺陷在于由于语言组织及其表达方式千变万化，就需要制定者有很丰富的语言知识，定义出许多模板，才能达到较好的效果。郭剑毅等人[8]利用改进的层叠条件随机场模型实现了旅游领域的命名实体识别(named entity recongnition)任务，取得了准确率、召回率和 F1 值均在 80%以上的效果；何宇等人[9]利用条件随机场作为抽取模型，选取词、词长、词性、依存关系、词典位置等作为特征模板，有效地抽取了新能源汽车领域术语。基于条件随机场的模型虽然能有效抽取领域术语，但是召回率不稳定，需要对文本标注和特征选取进行充分定义和选取。刘里等人 [10]提出一种基于术语长度和语法特征的统计领域术语抽取方法，在利用机器学习抽取候选术语时，加入基于术语长度和语法特征的约束规则，有效抽取了领域术语。综上分析，利用人工特征和领域知识的方法虽然取得了一定的识别效果，然而这种方法需要依据逻辑直觉人工定制大量特征，无法自动挖掘隐含特征，其识别性能很大程度上依赖所选取的特征的质量。

1　相关研究

在以上背景下，为解决新能源汽车领域术语抽取问题中存在的难点问题和现有方法中对于人工制定特征的过度依赖问题，本文从深度学习的角度提出一种全新的方法，它不需要详细地制定领域术语的特征表达，更具有实用性。近年来，利用基于神经网络的深度学习获取特征的方法在图像、语音以及自然语言处理领域取都备受关注。冯艳红等人[11]利用基于上下文的词向量和基于词的词向量提出了基于 BLSTM 的命名实体识别方法，并将领域知识嵌入模型的代价函数中，进一步增强函数的识别能力，取得了 95%的正确率；侯伟涛等人[12]使用双向 LSTM 神经网络学习文本的隐藏特征，解决了传统方法通用性不强以及无法有效利用后文信息的缺点，实现了医疗事件的识别研究；Raffel 等人[13]提出一种适用于前馈神经网络的简化的注意力模型，证明了 Attention 机制能够在文本较长的情况下有效解决信息丢失等长距离依赖问题。Yang 等人[14]提出的层次化注意网络有两个层次的注意机制，在单词和句子层次上应用，使它能够在构造文档表示时关注关键内容，充分说明了 Attention 机制能够给文本中的关键部分分配更多的注意力。张冲[15]设计了组合正逆序 attention-Based LSTM 模型，结合 Attention 机制和双向 LSTM 特点实现文本分类。注意力概率可以突出特定的单词对于整个句子的重要程度，并且考虑了更多的上下文语义关联。Li 等人[16]利用基于 CRF 的双向 LSTM 深层神经网络模型实现了对生物医学文本中的不规则实体的识别，正确率达到 81.09%。Gridach[17]利用 BLSTM 结合 CRF 模型实现的字符级神经网络完成了对生物医学命名实体的识别，在最终实现的系统上表现出了 90.27%的准确率。CRF 相对于其他模型可以更有效地关注上下文标注信息，所以结合 BLSTM 模型可以有效的改善实验结果。

在自然语言处理领域，基于深度学习的这类方法减少了人工定义特征和对领域知识的过度依赖，实现了端到端的命名实体识别模式。自动挖掘隐含特征可以有效的解决新词发现问题，所以本文利用深度学习的方法研究新能源汽车领域术语抽取问题。领域术语的抽取可以转换为序列标注问题。循环神经网络（recurrent neural network，RNN）是一种有效的解决序列标注问题的神经网络模型，但是 RNN 无法很好地处理自然语言处理中不可忽视的长距离依赖问题，并且其训练算法存在梯度消失或爆炸问题，而 LSTM 模型通过引入记忆单元和门限机制很好地解决了这个问题，但是 LSTM 只考虑上文信息，不考虑下文信息，双向的 LSTM（即 BLSTM）同时考虑上下文信息，对于本文的新能源汽车领域术语抽取问题具有极大意义。

本文提出 BLSTM-ATT-CRF+校正模型，首先对新能源汽车领域专利文本进行预处理，之后进行 Word Embedding 向量化，然后进入 BLSTM-ATT-CRF 模型进行标注。BLSTM-ATT-CRF 模型既考虑了上下文信息，有效解决了长距离依赖，又能通过 Attention 机制的加入计算注意力分配概率，有效防止信息的丢失，突出关键词的作用，同时与 CRF 结合，解决了 BLSTM 在处理输出标签时无法很好地处理有强烈依赖关系的数据的难题。在标注完成之后，建立基于词典与规则的校正模型，从而取得更好的标注效果。

2　新能源汽车领域术语抽取模型

本文首先对要处理的新能源汽车专利文本进行预处理，包括分词、词性标注、去停用词、标点过滤等。为降低运算复杂度和防止分词工具过度切分，本文添加停用词表和用户词典辅助分词系统进行分词。本文将新能源汽车领域术语的抽取转换为序列标注问题，提出 BLSTM-ATT-CRF 标注模型和基于词典与规则的校正模型。完整的标注流程如图 1 所示。

2.1　Word Embedding 向量化

文本预处理之后进行 Word Embedding 向量化[18,19]表示。Word Embedding 技术是一种采用机器学习将单词映射到实数低维向量的技术，可以避免传统词向量的维度过高或向量稀疏问题，还能提供含有语义信息的词向量。该过程如下：设样本句子 X 由 n 个词组成，X={t1,t2,...,tn}，其中第 t 个词 tt 为词的

one-hot 表示。词嵌入 xt 为

$$x_t = W^{emb} t_t \quad (1)$$

其中：$W^{emb} \in R^{d\times|v|}$，为 embedding 向量查询表，需要训练得到；$t_t \in R^{|v|}$，$x_t \in R_d$，d 为 embedding 向量维度，$|v|$为 one-hot 表示下词典的大小。

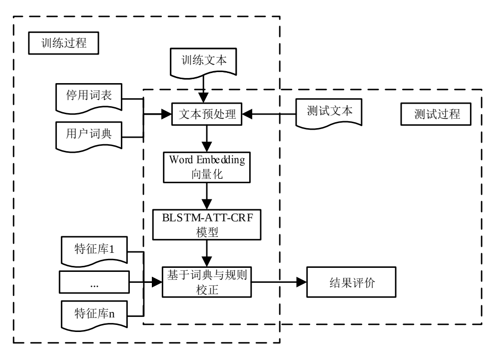


图 1　领域术语标注框架

2.2 BLSTM-ATT-CRF 模型概述

本文把术语抽取问题转换为序列标注问题。为了实现这一转换，需要对分词后的每一个词语进行标注。考虑新能源汽车专利文本中包含大量的英文术语，定义以下标注体系，即 R={B_cha，I_cha，E_cha，B_eng，I_eng，E_eng，O}，分别代表中文术语首部、中文术语中部、中文术语尾部、英文（缩写）术语首部、英文（缩写）术语中部、英文（缩写）术语尾部、其他。特别的，如果该术语只包含一个词，那么中英文只标注首部标签即可；如果该术语由中英文混合而成，标注之后由词语提取程序判断，连在一起的中英文则为一个领域术语。该标注体系的定义明确界定了词语边界，解决了领域术语中词长不定和中英文混合问题。

专利文本经过预处理和 Word Embedding 向量化之后，进入 BLSTM-ATT-CRF 模型进行训练。BLSTM-ATT-CRF 模型架构如图 2 所示。

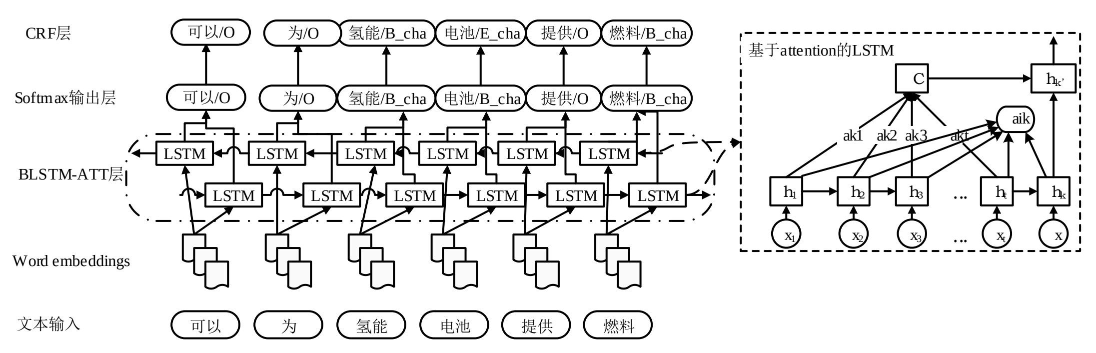


图 2　BLSTM-ATT-CRF 模型架构

2.2.1 BLSTM 模型

文献[20]中提出的 LSTM 引入记忆单元和门限机制，实现了对长距离信息的有效利用，解决了 RNN 模型中存在的梯度消失或者爆炸问题。但是 LSTM 只考虑文本的上文信息，不考虑下文信息，而对于本文的领域术语抽取问题，下文信息也很重要。Graves 等人[21]提出了双向的 LSTM（即 BLSTM）。BLSTM 有效地利用了文本序列的上下文信息，可以更多地挖掘隐含特征，有效解决新能源汽车领域术语抽取问题中的新词发现问题。BLSTM 结构如图 3 所示。

图 3 中 x_t 表示 BLSTM 模型在 t 时刻专利文本经过 Word Embedding 以后的向量化表示；$\overrightarrow{h_t}$ 是前向 LSTM 在 t 时刻的输出，$\overleftarrow{h_t}$ 为反向 LSTM 在 t 时刻的输出，所以 BLSTM 在 t 时刻的输出表示定义为 $h_t=[\overrightarrow{h_t},\overleftarrow{h_t}]$，就是直接拼接 $\overrightarrow{h_t}$ 和 $\overleftarrow{h_t}$。

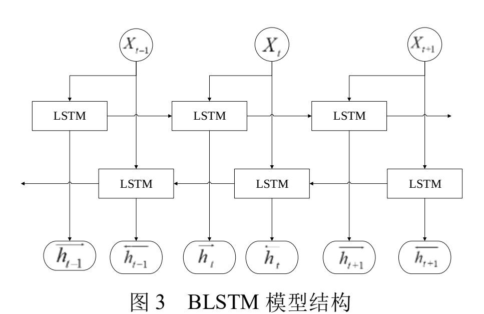


图 3　BLSTM 模型结构

2.2.2 Attention 机制

Attention 机制是一种模拟人脑注意力的机制，核心思想是借鉴了人脑在特定时刻对于事物的注意力会集中在某个关键点，而忽略其他非关键点，是一种人脑资源分配模型[22]。所以该机制的主要作用是对于关键词可以分配较多的注意力，而对于其他部分分配较少的注意力。将 Attention 机制的与 BLSTM 模型进行组合（见图 2 右侧基于 Attention 的 LSTM），有效突出关键词的作用。例如对于文本“汽车的制动是通过制动盘与制动钳或制动鼓与制动蹄之间的摩擦来实现的”和“制动能量回馈是提高汽车能量效率的一个非常重要的手段”，在不加入 Attention 机制的情况下，BLSTM 模型本身关注的是上下文信息，无法实现重点关注“制动”这个关键词，加入 Attention 之后，Attention 通过计算权重可以实现这一功能。文本预处理时进行了词性标注，考虑新能源汽车专利文本术语多数存在定中、动宾和主谓关系，在 BLSTM-ATT 层中同时把词性作为特征进行训练。本文 attention 机制的相关计算公式如下所示：

$$a_{ki} = \frac{\exp(e_{ki})}{\sum_{j=i}^{T}\exp(e_{kj})} \quad (2)$$

$$e_{ki} = v\tanh(Wh_k + Uh_i + b) \quad (3)$$

$$C = \sum_{i=1}^{T} \alpha_{ki} h_i \quad (4)$$

$$h_{k'} = H(C, h_k, X') \quad (5)$$

式（5）计算的是注意力概率分布的语义编码，a_{ki} 计算的是节点 i 对于节点 k 的注意力概率权重。T 为输入序列的元素的数目。V、W、U 为权重矩阵，本文中 attention 机制的输入为上文 BLSTM 的输出；h_i 为 BLSTM 模型中前向输出 $\overrightarrow{h_t}$；h_k 为 BLSTM 模型中反向输出 $\overleftarrow{h_t}$。BLSTM 输出的所有结果都进入 attention 进行计算。C 是语义编码。$h_{k'}$ 就是最终的特征向量，该特征向量表现为突出关键词的语义信息。

2.2.3 BLSTM-ATT-CRF 模型

条件随机场模型（conditional random fields，CRF）是 Lafferty 等人[23]提出的一种无向图的模型，在中文分词、命名实体识别、歧义消解等汉语自然语言处理任务中都有应用，并有着良好表现[24]。在基于 Attention 机制的 BLSTM 模型中引入 CRF 模型，使得模型在结合上下文信息的同时可以有效地考虑输出标签前后的依赖关系。实际效果中 BLSTM-ATT 模型和在加入 CRF 之后的 BLSTM-ATT-CRF 模型表现如表 1 所示。原因就是 CRF 模型可以有效考虑标签 B_cha 与 E_cha 之间的依赖关系，所以可以正确识别出此处“功率分析仪”是一个完整术语，而 BLSTM-ATT 模型无法做到这一点，所以会错误的将“功率分析仪”划分为两个术语。因此，CRF 模型的加入可以有效解决领域术语多为嵌套和复合结构的识别问题。

具体做法就是在 BLSTM 的 softmax 输出层之后加入 CRF 层，引入状态转移矩阵 A 作为 CRF 层的参数，设矩阵 L 为 BLSTM 的输出，其中 $A_{i,j}$ 表示时间顺序上从第 i 个状态转移到第 j 个状态的概率；$L_{i,j}$ 表示观察序列中第 i 个词被标注为第 j 个标注的概率。本文采用最大似然估计作为代价函数，采用维特比算法解码。观察序列 X 的待预测标注序列 Y=（$y_1,y_2,\dots,y_n$）的输出计算公式为

$$s(X,Y) = \sum_{i=1}^{n} (A_{yi,yi+1} + L_{i,yi}) \quad (6)$$

$$\log L(y|x) = s(X,Y) - \log \sum_{y'} \exp(s(X,Y')) \quad (7)$$

表 1　BLSTM-ATT 模型和 BLSTM-ATT-CRF 模型标注对比

模型	标注结果
BLSTM-ATT	通过/O 功率/B_cha 分析仪/B_cha 采集/O 电机/B_cha 运行/O 的/O 输入/O 电流/B_cha 和/O 输入/O 电压/B_cha
BLSTM-ATT-CRF	通过/O 功率/B_cha 分析仪/E_cha 采集/O 电机/B_cha 运行/O 的/O 输入/O 电流/B_cha 和/O 输入/O 电压/B_cha

2.3　基于词典与规则的校正

经过对新能源汽车的领域特征及语言特征进行分析统计，其领域术语构成存在特定规律。而专利文本中新能源汽车领域术语以名词结尾的占 86.47%[22]，其中又包含一些常用关键词，如器、车、机等；新能源汽车领域术语首词与中心词之间大多为定中关系、主谓关系和动宾关系，约占 78%[25]。所以本文最后采用基于词典与规则相结合的方法对 BLSTM-ATT-CRF 模型的识别结果进行校正，以提高抽取结果的正确率。对新能源汽车领域术语进行校正的对象是新能源汽车领域术语的中文表述和别名。新能源汽车领域术语大多为名词短语。分析发现虽然构成新能源汽车领域术语的词性组合有多种模式，但是词性主要为名词、动名词和形容词这三种为主。据此，本文通过分析《GB/T 19596-2004 电动汽车术语》、《GB/T 28382-2012 纯电动乘用车技术条件》、《GB/T24548-2009 燃料电池电动汽车术语》和《GB/T 20042.1 质子交换膜燃料电池术语》等文献中新能源汽车领域中所包含的术语特征，人工建立新能源汽车领域术语特征词库，如表 2 所示。通过总结文献中新能源汽车领域术语构词规律，建立新能源汽车领域术语构词特征库，如表 3 所示。并制定相应规则进行判断：

表 2　新能源汽车领域术语特征词库示例

特征词类	示例
常用名词（A）	活塞、连杆、轴承、涡轮、…
常用动词(B)	供油、喷油、喷气、进气、…
常用词缀（尾）（C）	器、车、机、环、缸…
常用形容词（D）	管式、粘结式、耐油、抗静电、...
其他词类	天干（E）、希腊词母（F）、汉文数词（G）、罗马数词（H）

表 3　新能源汽车领域术语构词特征库

构词特征	示例
n	控制器、扶手、车厢、...
n+n	发动机舱、蓄电池箱、车身附件、发动机罩、…
n+n+n	泡沫塑料软垫、乳胶棕丝软垫、轮缘端部半径、座椅中心平面、…
n+vn+n	安全门开启角、螺栓孔分布圆直径、带束斜交轮胎、花纹加强筋…
n+n+vn+n	胎面花纹展开图、轮胎气门嘴（孔）位置、机油油量报警传感器...
n+n+n+n	航空轮胎圆形截面、实心轮胎基部宽度、...
ad+n	厢式货箱、单人座椅、对开式车轮、嵌入式头枕、...
ad+n+n	管式自行车轮胎、粘结式实心轮胎、非粘结式实心轮胎、...
n+ad+n	双胎最小间距、电磁振动式调节器、机械啮合式起动机、...

规则 1　包含表 1 中标号 D-H 的新能源汽车领域术语在记录中的比例不足 1%。若模型求出的新能源汽车领域术语包含标号 D-H 中的内容，则移到下一个词，继续判断是否包含。若包含继续移动，直到不包含或者包含词库中其他词为止。

规则 2　被标注的新能源汽车领域术语如果符合表 2 中的其中一种形式，并且由表 1 中词库中已经存在的词语组合得到，那么该标注序列直接标记为有效的序列。

规则 3　在表 1 特征词库的基础上，提出一种文本术语匹配算法。算法流程如图 4 所示。当库中词条能与待校正样本完全匹配时，无论模型求得的结果如何，都将该词组进行标注作为有效序列。

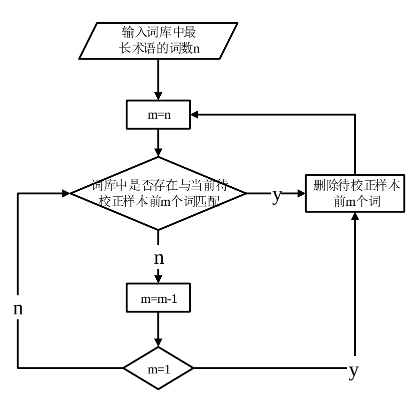


图 4　文本术语匹配算法流程

3　实验设计与结果分析

3.1　实验语料及评价标准

目前还没有出现公认度较高的面向新能源汽车领域的语料，所以本文实验采用的数据是从专利网上下载的新能源汽车领域专利文本。由于实验最终获取的是专利文本中的领域术语，而领域术语分布在整篇专利的中的各个部分，为了保证实验的通用性，实验采用的是整篇的专利文本，本实验共标注专利文本 1 126 篇，人工标注结果已经在 CAI 创新工具中得到验证，其中 800 篇用于训练过程，326 篇用于测试过程。标注样例如表 4 所示。两个过程文本数量及领域术语标注数量如表 5 所示。

表 4　新能源汽车领域术语文本标注样例

公开 /v	一 /m	种 /q	涡旋 /n	电动 /b	真空泵 /n	， /wd	包括 /v
O	O	O	B-cha	I_cha	E_cha	O	O
外壳体 /n	、 /wn	动 /v	涡旋 /n	盘 /qv	。/wj	VCU /n	采集 /vn
B-cha	O	O	O	O	O	B-eng	B-cha
传感器 /n	信号 /n	。/wj	从/p	高压/n	到/v	低压/n	的/u
I_cha	E_cha	O	O	B-cha	O	B-cha	O
DC/ws	//w	DC/ws	转换器/n	。/wj			
B-eng	I-eng	E-eng	B-cha	O			

表 5　训练过程和测试过程数据表

	专利/篇	中文术语/个	英文术语/个
训练集	800	43 127	315
测试集	326	18 457	105

本文实验为了减少外在人为因素的影响，采用三倍交叉验证方式。本文采用式（8）～（10）指标衡量实验结果：准确率 P，召回率 R，F1 值。

$$P=\frac{识别出的正确的术语个数(RN)}{识别出的术语个数(STN)}\times 100\% \tag{8}$$

$$R=\frac{识别出正确术语的个数(RN)}{所有标注的术语的个数(TN)}\times 100\% \tag{9}$$

$$F1=\frac{2PR}{P+R} \tag{10}$$

3.2　实验设计

本文使用中科院研制的汉语词法分析系统 ICTCLAS 对语料进行分词。为了获得高质量的 embedding 向量查询表，本文首先提取《GB/T 19596-2004 电动汽车术语》、《GB/T 28382-2012 纯电动乘用车技术条件》、《GB/T24548-2009 燃料电池电动汽车术语》和《GB/T 20042.1 质子交换膜燃料电池术语》等文献中新能源汽车领域中所包含的术语词条，然后利用 word2vec 工具中的 Skip-gram 模型进行训练得到。

BLSTM-ATT-CRF 模型的运行环境为 64 位 Windows 7 操作系统，运行内存为 8 GB。模型使用了 keras 与 theano 的集成框架实现，实现语言为 Python；框架安转用到第三方平台为 Anaconda2，并使用反向传播算法（BPTT）[26]进行训练。CRF 模型使用的是 CRF++ 0.58 工具包，其中的模型参数值如-c、-f 等是根据人工经验设定。该模型是否能够取得较好的识别效果与其参数具有密切关系。其中涉及到的参数有 embedding 向量维度、学习率、隐藏层单元数量以及 dropout[27]值。本文对这四种参数进行实验，得到实验效果最好的值作为模型的参数。实验表明，各参数均存在局部最优值，当 embedding 向量维度为 200、学习率为 0.02、隐藏层单元数量为 200 以及 Dropout 值为 0.1 时，模型的实验效果达到最佳。除此之外，各参数对实验效果的影响也不尽相同，在此实验中，embedding 向量维数对实验效果影响最大，隐藏层单元数量对实验效果的影响较小。

为了检验本文模型对于新能源汽车领域术语标注的效果，本文设计了多组实验来进行对比分析，共包含八组实验：传统的 LSTM 模型实验、传统的 RNN 模型实验、传统的 CRF 模型实验、双向的 LSTM 模型实验（BLSTM）、BLSTM-CRF 模型实验、基于 Attention 机制的 BLSTM 模型实验（BLSTM-ATT）、，BLSTM-ATT-CRF 模型实验和 BLSTM-ATT-CRF+校正模型实验。特征选择如表 6 所示。其他模型所用参数均与上文得出的参数相同。

表 6　CRF 特征选择

1	Cur_Char	当前词
2	Cur_Char_Label	当前词的标注标签
3	Cur_Char_Part	当前词的词性
4	Cur_OnePre_Char_Label	当前词的前面第一个词的标注标签
5	Cur_TwoPre_Char_Label	当前词的前面第二个词的标注标签
6	Cur_OneAfter_Char_Label	当前词的后面第一个词的标注标签
7	Cur_TwoAfter_Char_Label	当前词的后面第二个词的标注标签
8	Cur_OnePre_Char_Part	当前词的前面第一个词的词性
9	Cur_TwoPre_Char_Part	当前词的前面第二个词的词性
10	Cur_OneAfter_Char_Part	当前词的后面第一个词的词性
11	Cur_TwoAfter_Char_Part	当前词的后面第二个词的词性

3.3　实验结果及分析

经上述八种模型实验后，实验结果如表 7 所示。从表 7 所展示的实验数据可以看出，本文设计的 BLSTM-ATT-CRF+校正模型能够有效地提高新能源汽车领域术语抽取的效果。通过以上实验可以看出，未经过改进的模型 1、2、3 虽然在一定程度上能够解决本文的领域术语抽取问题，但是整体来看取得的效果不佳。虽然 LSTM 解决了 RNN 梯度消失的问题，但是 LSTM 仅考虑上文信息，也不能取得较好的实验效果。实验 4 中 BLSTM 模型解决了 LSTM 的这一问题，综合考率文本的上下文信息，使得抽取的效果有了较明显的提升，准确率达到了 80.01%。实验 6 比实验 4 的正确率提高了 1.57%，F1 值提高了 1.29%。其主要原因是前者在合并输出向量时候 Attention 机制为每个输出向量赋予了不同的权值，使模型将注意力集中在更重要的向量上，从而降低了无关向量的作用。这个模型能够更好地表征文本，突出关键词的作用。通过实验 6 和 7 的对比可以看出，CRF 模型的引入对于新能源汽车领域术语抽取具有一定的意义。这是因为 BLSTM-ATT-CRF 模型在考虑上下文信息的同时由于 CRF 特征模板的合理制定还考虑了句子前后的标签信息，所以该模型能够取得不错的效果。通过实验 7 和 8 的对比可以看出，引入校正模块能够明显提升正确率和召回率。这是因为基于词典与规则校正模型是基于对句子结构的深入分析，并且考虑专利文本的表达习惯而提出的，其更能在尊重句子原义的基础上进行判断分析。综上所述，本文设计的 BLSTM-ATT-CRF+校正模型可以取得比一般深度学习模型更好的实验效果。

表 7　实验结果

实验标号	模型名称	指标		
		P_准确率/%	R_召回率/%	F1 值/%
1	LSTM	77.26	75.66	76.45
2	RNN	75.55	72.57	74.03
3	CRF	73.53	68.42	70.88
4	BLSTM	80.01	78.89	79.44
5	BLSTM-CRF	82.24	79.09	80.63
6	BLSTM-ATT	81.58	79.91	80.73
7	BLSTM-ATT-CRF	84.27	81.99	83.11
8	BLSTM-ATT-CRF+校正	86.62	85.07	85.83

4　结束语

综上所述，本文提出一种面向新能源汽车领域专利文本的领域术语抽取方法。本文首先建立了 BLSTM-ATT-CRF 模型，该模型解决了文本标注过程中存在的多种共性问题，所以该模型同样适用于其他领域的领域术语抽取问题。但是为了提高本文新能源汽车领域术语标注的准确率以及 F1 值，在经过 BLSTM-ATT-CRF 模型标注的基础之上，文本深入挖掘领域术语的句子构成和专利文本的表达特征，进一步制定了基于词典和规则的校正模型。经过对比实验表明，本文模型具有较好的准确性和鲁棒性。接下来将在继续增加语料的基础上对方法继续优化，对词典和校正规则进行进一步扩充，使抽取结果更加严谨而有效，使得模型具有更好的泛化性。

参考文献：

[1] Zhu Xiaojin. Semi-supervised learning literature survey, TR-1530 [R]. [S. l.] : University of Wisconsin-Madison. 2008.

[2] 王密平. 汉语专利术语抽取及应用研究 [D]. 南京: 南京大学, 2017.

[3] 樊梦佳, 段东圣, 杜翠兰, 等. 统计与规则相融合的领域术语抽取算法 [J]. 计算机应用研究, 2016, 33 (8): 2282-2285, 2306.

[4] 葛煦, 卢宝华, 杨湘华, 等. 谈高校科技发展中专利文献的利用 [J]. 技术与创新管理, 2005, 26 (1): 68-70.

[5] 贾志琦, 邵曰剑. 有效利用专利文献提高企业技术创新能力 [J]. 山西科技, 2008 (1): 91-93

[6] 王密平, 王昊, 邓三鸿, 等. 基于 CRF 的冶金领域中文专利术语抽取研究 [J]. 现代图书情报技术, 2016 (6): 28-36

[7] 周浪, 史树敏, 冯冲, 等. 基于多策略融合的中文术语抽取方法 [J]. 情报学报, 2010, 29 (3): 460-467.

[8] 郭剑毅, 薛征山, 余正涛, 等. 基于层叠条件随机场的旅游领域命名实体识别 [J]. 中文信息学报, 2009, 23 (5): 47-52.

[9] 何宇, 吕学强, 徐丽萍. 新能源汽车领域中文术语抽取方法 [J]. 现代图书情报技术, 2015 (10): 88-94.

[10] 刘里, 肖迎元. 基于术语长度和语法特征的统计领域术语抽取 [J]. 哈尔滨工程大学学报, 2017 (9): 1437-1443.

[11] 冯艳红, 于红, 孙庚, 等. 基于 BLSTM 的命名实体识别方法 [J//OL]. 计算机科学, 2018 (2): (2017-05-16) .

[12] 侯伟涛, 姬东鸿. 基于 Bi-LSTM 的医疗事件识别研究 [J//OL]. 计算机应用研究, 2018, 35 (7): 1-2 (2017-07-27) .

[13] Raffel C, Ellis D P W. Feed-forward networks with attention can solve some long-term memory problems [C]// Proc of ICLR 2016 Workshop Submissionreaders. 2016.

[14] Yang Z, Yang D, Dyer C, *et al.* Hierarchical attention networks for document classification [C]// Proc of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2017: 1480-1489.

[15] 张冲. 基于 Attention-Based LSTM 模型的文本分类技术的研究 [D]. 南京: 南京大学, 2016.

[16] Li Fei, Zhang Meishan, Tian Bo, *et al.* Recognizing irregular entities in biomedical text via deep neural networks [C]// Proc of Pattern Recognition Letters. 2017.

[17] Mourad Gridach. Character-level neural network for biomedical named entity recognition [J]. Journal of Biomedical Informatics, 2017, 70.

[18] Mikolov T, Sutskever I, Chen K, *et al.* Distributed representations of words and phrases and their compositionality [C]// Advances in Neural Information

Processing Systems. 2013: 3111-3119.

[19] 孟欣，左万利. 基于 word embedding 的短文本特征扩展与分类 [J]. 小型微型计算机系统，2017，38 (8): 1712-1717.

[20] Jozefowicz R, Zaremba W, Sutskever I. An empirical exploration of recurrent network architectures [C]// Proc of International Conference on Machine Learning. 2015: 2342-2350

[21] Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures [J]. Nrural Networks, 2005, 18 (5): 602-610.

[22] 张冲. 基于 Attention-based LSTM 模型的文本分类技术的研究 [D]. 南京：南京大学, 2016.

[23] Lafferty J, McCallum A, Pereira F. Conditional random fields: probabilistic models for segmenting and labeling sequence data [C]// Proc of the 18th International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers, 2001: 282-289

[24] 郑敏洁，雷志城，廖祥文，等. 中文句子评价对象抽取的特征分析研究 [J]. 福州大学学报：自然科学版, 2012, 40 (5): 584-590.

[25] 何宇，吕学强，徐丽萍. 新能源汽车领域中文术语抽取方法 [J]. 现代图书情报技术, 2015 (10): 88-94.

[26] Werbos P J. Backpropagation through time: what it does and how to do it [J]. Proceedings of the IEEE, 1990, 78 (10): 1550-1560.

[27] Hinton G E, Srivastava N, Krizhevsky A, *et al.* Improving neural networks by preventing co-adaptation of feature detectors [J]. Computer Science, 2012, 3 (4): 212-223.